ウェブ上のアンカーテキスト間の類似尺度

林 裕史　　岡部 正幸　　梅村 恭司
豊橋技術科学大学 応用数理ネットワーク研究室
hayashi@ss.cs.tut.ac.jp, okabe@imc.tut.ac.jp, umemura@tut.jp

1　はじめに

ウェブ上のアンカーテキストは，ユーザがリンク先コンテンツの価値を判断するための重要な要素であり，リンク先コンテンツの概要を示す文字列として記述される場合が多い．先行研究 [1] においてはアンカーテキストを利用したことで Web 検索の精度が向上すると報告されている．そこで，アンカーテキスト間の類似度からリンク先コンテンツ間の類似度を判断することができれば，システム上有用であると考えた．本研究ではウェブ上のアンカーテキスト間の類似尺度を提案し，またその類似尺度を評価することを研究目的とする．

多くの場合，共通したコンテンツを指すアンカーテキストのペアの間にも同義語や表記ゆれによる差異が存在し，全く同じ文字列になることは少ない．例えば商店ウェブサイトや学校ウェブサイトには所在地周辺の地図や経路情報を掲載したコンテンツがよく見られるが，それらを指すアンカーテキストには「交通案内」「交通アクセス」「アクセスマップ」など，複数のパターンが考えられる．そういったアンカーテキストのペアに適した類似尺度を見つけることは興味深い問題である．この問題に対し本研究では，同義語辞書と編集距離を用いたアプローチを行った．

同義語辞書は人手により作成し，アンカーテキストに対して最長一致法による語句の置換を行うことで同義語の統一を図った．編集距離は二つの文字列間の相違度を表す尺度の一つとして知られており，文字列間の距離として類似度を定量的に表現する基準として利用した．この手法は，例えば「ユーザ」「ユーザー」といった表記ゆれによりマッチングし難い文字列のペア間に対して距離の増加を少量に抑え，類似度を高く算出するためにも有効であると考えた．

本研究では，まず同義語辞書の有無による効果を評価するため，同義語の統一を行わず編集距離のみを用いる類似尺度と，同義語の統一を行った後に編集距離を用いる類似尺度とについて，それぞれを用いて正解集合の各要素に対する距離を求め，比較評価を行った．評価のために本研究で整備した正解集合は，複数の高専ウェブサイトから計 2,746 種のアンカーテキストを取得し，その中で高専ウェブサイトに頻出すると判断したコンテンツを指すアンカーテキストを全体集合として，類似コンテンツを指すアンカーテキストのペアを人手で選出して作成した．

評価の結果，同義語の統一を行わなかった類似尺度に比べ，同義語の統一を行った類似尺度が全体的に小さい距離を示した．また正解集合中の多くのペア間では距離が小さく算出されたことから，類似コンテンツを指すアンカーテキスト間の類似尺度として一定の効果があったことを報告する．

2　辞書による同義語統一

人手で作成した同義語辞書を用いて，最長一致法によりアンカーテキストの文字列を置換することで同義語の統一を図った．定義した同義語グループをリスト 1 に示す．

FAQ　ＦＡＱ　Q&A　Ｑ＆Ａ　よくある質問　質問集　問答集
ロボコン　ロボットコンテスト　Robot Contest
技術　テクノロジー　Technology
情報　インフォメーション　インフォ　Information　Info
交通案内　アクセスマップ　交通アクセス　周辺地図　周辺マップ
資料　パンフレット　小冊子　冊子
プロフィール　Profile　略歴
リンク　Link
メンバー　Member
案内　紹介
の方へ　の皆さんへ　の皆様へ　の皆さまへ
在校生　在学生
目標　方針

リスト 1: 辞書の同義語グループ定義

この辞書を用いてアンカーテキストに対して以下の置換処理を行うことで同義語統一処理を実装した．

1. アンカーテキスト中の置換対象となる語句の選択は，辞書中に記載されている各語句についてアン

カーテキストを走査し，最も先頭に近い位置に現れる語句のうち，最も語長の長いものとする．

2. 置換先となる語句の選択は，同義語グループの中で最も語長の短い語句とする．ただし最も語長の短い語句が複数存在する場合は，最も手前に記載されているものとする．

3. アンカーテキストの末尾に到達するまで，置換された部分以降のアンカーテキストについて 1〜2 の処理を繰り返す．

例えば「ロボットコンテストに関する Q&A」といったアンカーテキストがあった場合，以下のように同義語統一処理を行う．

1. 辞書中に記載されている語句のうち，最も先頭に近い「ロボットコンテスト」が置換対象となる．

2. グループ内で最も短い語句である「ロボコン」に置換される．

3. 残りの「に関する Q&A」に対して，「Q&A」が唯一ヒットし置換対象となる．

4. グループ内で最も短い語句は「FAQ」「ＦＡＱ」「Q&A」「Ｑ＆Ａ」となり，最も手前に記載されている「FAQ」に置換される．

5. アンカーテキストの末尾に到達し，「ロボコンに関する FAQ」が出力となる．

3　編集距離

編集距離 [2] はレーヴェンシュタイン距離とも呼ばれ，一方の文字列を他方の文字列へと変換するために必要な文字の操作のコストより算出される．この距離が小さいほど，それら二つの文字列間の類似度が高いと言える．操作の種類や文字種に応じて編集距離に重みを付ける場合もあり，その場合は重み付き編集距離とも呼ばれる．本研究では文字の持つ意味を重視するため，表 1 のように文字種ごとに重みを設定した．

表 1: 文字種による距離の重み

文字種	重み
アルファベット・記号	0.8
ひらがな・カタカナ	1.0
漢字	2.0

例えば「交通案内」と「交通アクセス」の間の距離は，図 1 に示す手順で求められる．

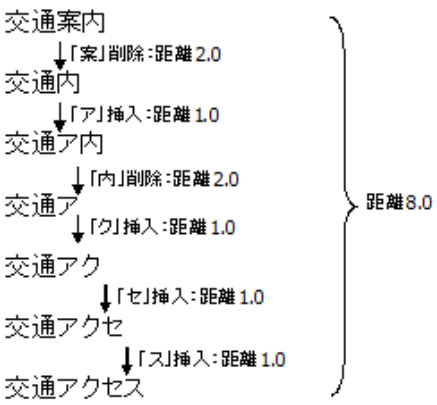

図 1: 編集距離算出手順

最短距離を算出するアルゴリズムは図 2 に示す編集グラフの概念を用いて実装した．

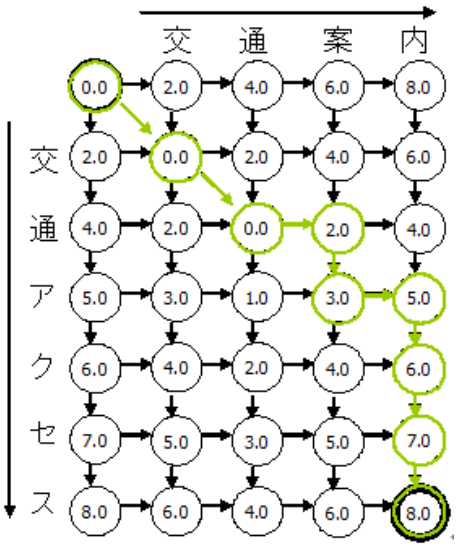

図 2: 編集グラフの概念図

編集グラフは，文字列変換作業の経過をノードとして表した二次元グラフである．左上端のノードが変換前，右下端のノードが変換完了後の文字列を表し，各ノードへの到達コストを順に計算していくことで最短距離を求めた．

グラフにおける下方向，右方向への移動は文字の操作である挿入，削除に相当する．あるノードから見て下方向の文字と右方向の文字が同じである場合に限り，文字の変更が不要であるため，コスト 0 で斜め右下方向への移動が可能とする．

最上段および最左列の各ノードの到達経路は一通りしかないため，到達コストは簡単に求まる．その他の各ノードの到達コストは，左方向・上方向・斜め左上方向のうち，いずれから来た場合に距離が最も小さく

なるかを計算することで求まる．

以上の操作を順に繰り返し右下端のノードまでグラフを埋めることで最短距離を求めた．

4　同義語辞書と編集距離による類似尺度の評価

評価のために作成した正解集合は，クローリングにより豊田高専ウェブサイトから 1,589 種，大島商船ウェブサイトから 350 種，熊本高専ウェブサイトから 807 種，計 2,746 種のアンカーテキストを収集し，その中で高専ウェブサイトに頻出すると判断したコンテンツを指すアンカーテキストを全体集合として，類似コンテンツを指すアンカーテキストのペア 149 組を人手で選出した．収集したアンカーテキストのリストを付録 A-1 に，選出した正解集合のリストを付録 A-2 に示す．また，選出したペアの例をリスト 2 に示す．

「教育課程(学科・専攻科)」	「学科・専攻科紹介」
「学校紹介」	「学校案内」
「中学生の皆さんへ」	「受験生の方へ」
「豊田高専の沿革」	「沿革」
「校長のメッセージ」	「校長挨拶」

リスト 2: 正解集合のペアの例

まず同義語辞書の有無による効果を確認するために，同義語辞書による統一を行わない類似尺度と統一を行う類似尺度とを比較した．それぞれの尺度により正解集合の各ペア間の距離を算出した結果を表 2 に示す．

表 2: 正解集合の距離分布表

距離	辞書あり	辞書なし
0	37	49
1〜4	14	12
5〜8	28	25
9 以上	70	63

同義語辞書による統一を行わない場合に比べ，統一を行った場合には距離 0 となった組数が 12 増加し，また距離が 1 以上となった組数は全ての領域で減少した．すなわち，同義語の統一によって全体としてペア間の距離が小さく算出されたといえる．ただし，正解集合であるにも関わらず距離が大きく算出されたペアが全体の 40%程度あった．期待に反して距離が大きく算出されたペアの例をリスト 3 に示す．

これらのペアについて距離が大きく算出された原因については以下のように考えた．

ペア		距離	
校長のメッセージ	校長挨拶	9.8	・・・A)
認証評価	機関別認証評価の結果	11.0	・・・B)
教育目標	理念・目的・育成する人物像	25.6	・・・C)

リスト 3: 距離が大きく算出されたペアの例

A) 「メッセージ」と「挨拶」といった，置換が妥当かどうかの判断が難しい単語が含まれており，今回用いた同義語辞書では統一が行えなかった．

B) 一方のアンカーテキストは全体が一致しているが，他方のアンカーテキストの文字列長が長いことにより不一致部分が増え，距離が大きく算出された．

C) 語句を言い換えた表現に対して今回用いた同義語辞書では統一が行えなかった．また，語句が並列表現された文字列に対して単に編集距離で評価するのは適切でなかった．

5　辞書の有無による平均編集距離の差

同義語辞書の有無による効果がどの程度良い結果であったのかを評価するため，全体集合のペア 2,205,605 組と正解集合のペア 149 組のそれぞれに対して，同義語辞書の有無による平均距離の変化を測定し，比較した．結果を表 3 に示す．

表 3: 平均距離の変化

	全体集合	正解集合
辞書なし	48.8	9.4
辞書あり	48.7	8.3
変化率	-0.2%	-11.6%

同義語辞書の有無による全体集合の平均距離の変化率は-0.2%であったのに対し，正解集合の平均距離の変化率は-11.6%と，全体集合の場合に比べ 58 倍程度変化した．このことから，同義語の統一は特に正解集合のペアの距離を小さくする効果があることがわかった．

6　まとめ

ウェブ上で類似度の高いコンテンツを指すアンカーテキスト間の類似尺度として，同義語辞書と編集距離を用いた手法を提案し，評価した．評価に用いた正解

集合は，複数の高専ウェブサイトから収集した類似コンテンツを指すアンカーテキストのペア 149 組を要素として作成した．

まず同義語辞書の有無による効果を確認するために，辞書による同義語の統一を行わない類似尺度と統一を行う類似尺度とを比較した．評価の結果，同義語の統一を行わない場合に比べ，統一を行った場合には全体としてペア間の距離が小さく算出された．ただし，正解集合のペアであるにも関わらずうまくマッチングできず，距離が大きく算出されたものも多かった．

また，全体集合のペア 2,205,605 組と正解集合のペア 149 組のそれぞれの同義語辞書の有無による平均距離の変化率を比較した結果，同義語の統一は特に正解集合のペアの距離を小さくする効果があることがわかった．

7　おわりに

提案した類似尺度の効果をより正確に評価するためには，不正解集合に対しても同様の評価実験を行い，その結果との比較を行う必要がある．不正解集合はどのようなペアを要素として構成するのが妥当であるかという問題も含め，本研究における今後の課題である．

参考文献

[1] Yinghui Xu, Kyoji Umemura, “ Web Searching Using Term Entropy on Virtual Document and Query Independent Importance in NTCIR-4 Web Task ” NTCIR Workshop 4 Meeting, Vol.1, pp.1-8, 2004

[2] 編集距離アルゴリズムを使って文字列を変換する - japan.internet.com
http://japan.internet.com/developer/20100219/26.html